(19) JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11) Publication number: 08166965 A

(43) Date of publication of application: 25 . 06 . 96

(51) Int. Cl G06F 17/30
G06F 17/27

(21) Application number: 06310875

(22) Date of filing: 14 . 12 . 94

(71) Applicant: NIPPON TELEGR & TELEPH CORP <NTT>

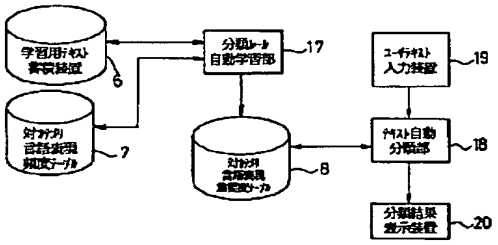
(72) Inventor: SUNABA RINTAROU

(54) METHOD FOR AUTOMATICALLY CLASSIFYING JAPANESE TEXT

(57) Abstract:

PURPOSE: To automatically classify a newly inputted Japanese text by learning appearance frequency information of a word (a noun, a verb, an adjective and an adverb) being intrinsic to a category and of language expression being equal to a modifier and a word to be modified in a text database which is previously classified into several categories.

CONSTITUTION: An automatic classification rule learning part 17 accesses to a learning text storing device 6 and executes learning from the classified text so that anti-category language expression importance degree tables 7 and 8 are generated. Then, an automatic text classifying part 18 accesses to the anti- category language expression importance degree table 8 as against the text inputted from a user text input device 19 and a classified result is outputted from a classification result display device 20.



(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平8－166965

(43)公開日　平成８年(1996)６月25日

(51)Int.Cl.[6]	識別記号	庁内整理番号	ＦＩ			技術表示箇所
Ｇ０６Ｆ 17/30						
17/27						
		9194－5L	Ｇ０６Ｆ	15/ 403	３５０　Ｚ	
		8420－5L		15/ 38	Ｄ	
		9194－5L		15/ 403	３４０　Ｂ	

審査請求　未請求　請求項の数２　ＯＬ　（全９頁）

(21)出願番号　特願平6－310875

(22)出願日　平成６年(1994)12月14日

(71)出願人　000004226
日本電信電話株式会社
東京都新宿区西新宿三丁目19番２号

(72)発明者　砂場　倫太郎
東京都千代田区内幸町１丁目１番６号　日本電信電話株式会社内

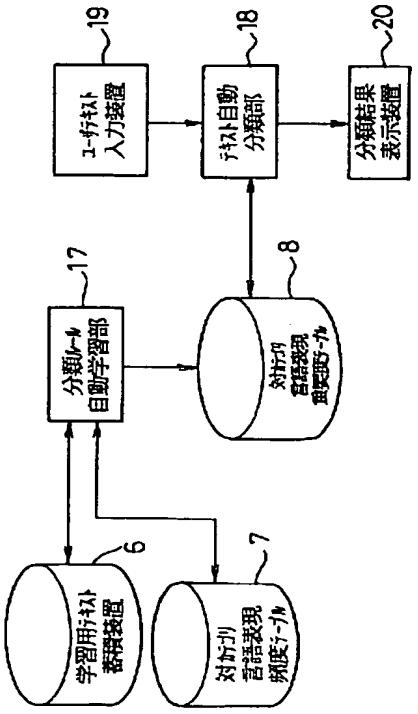
(74)代理人　弁理士　三好　秀和　（外１名）

(54)【発明の名称】　日本語テキスト自動分類方法

(57)【要約】

【目的】　予めいくつかのカテゴリに分類されたテキストデータベースに対して、カテゴリ固有の単語（名詞、動詞、形容詞、形容動詞）および修飾語・被修飾語対等の言語表現の出現頻度情報を学習することによって新規に入力された日本語テキストを自動的に分類する日本語テキスト自動分類方法を提供する。

【構成】　分類ルール自動学習部１７が学習用テキスト蓄積装置６をアクセスして分類済みのテキストから学習することにより対カテゴリ言語表現重要度テーブル７および対カテゴリ言語表現重要度テーブル８を作成し、ユーザテキスト入力装置１９から入力されたテキストに対してテキスト自動分類部１８が対カテゴリ言語表現重要度テーブル８をアクセスして分類した結果を分類結果表示装置２０から出力する。



1

【特許請求の範囲】

【請求項１】　日本語のテキストに対して単語および単語の組の頻度を特徴として抽出し、テキストの分類を行う日本語テキスト自動分類方法であって、

分類済みテキストアクセス工程にて学習用テキスト蓄積装置に蓄積されている分類ルール抽出のためのテキストをカテゴリ毎にアクセスし、

言語表現頻度解析工程にて入力テキスト中の名詞、動詞、形容詞、形容動詞、および修飾語・被修飾語対といった言語表現の出現頻度を計測し、

対カテゴリ言語表現頻度テーブル作成工程にて各カテゴリ毎の言語表現の出現頻度の蓄積テーブルを作成し、頻度計測終了判定の後に、

対カテゴリ言語表現重要度テーブル作成工程にて、各カテゴリ毎の言語表現の出現頻度を正規化した値の蓄積テーブルを作成する分類ルール自動学習工程と、

新規テキスト入力工程にてカテゴリ判定のための新規テキストを入力し、言語表現類似度判定工程にて新規のテキストに出現する言語表現の頻度と、カテゴリ毎の言語表現重要度との類似度を計算した後、該新規テキストのカテゴリを判定し、

分類結果出力工程にて前記新規テキストのカテゴリ判定結果を出力するテキスト自動分類工程とを備えたことを特徴とする日本語テキスト自動分類方法。

【請求項２】　前記分類ルール自動学習工程内の言語表現頻度解析において、入力テキストを単語に分割し、名詞、動詞、形容詞、形容動詞といった自立語をラベルし、形態素解析を行う工程と、

形態素解析の結果から、修飾語と被修飾語の対を抽出し、修飾語・被修飾語解析を行う工程と、

形態素解析と修飾語・被修飾語解析の結果から言語表現のリストを作成する言語表現抽出工程と、

入力テキスト中の言語表現の出現頻度を計測する言語表現出現頻度測定工程とを有することを特徴とする請求項１記載の日本語テキスト自動分類方法。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は、日本語テキスト処理装置などにおいてテキストに出現する単語（名詞、形容詞、動詞、形容動詞）、更に単語の組に注目し、予め分類されたテキストデータベース中の単語および単語の組の頻度を特徴として抽出し、新規のテキストの分類を行う日本語テキスト自動分類方法に関する。

【０００２】

【従来の技術】従来のテキスト分類方法には各種のものが存在する。例えば、従来のテキスト分類方法では、テキストの分類のための手がかりとして、主としてテキスト中の人名、学術用語、製品名といった名詞をキーワードに用いるものがあるが、この場合にはカテゴリを特徴付けるキーワードやキーワードの組合せパターンを人手で作成し、その条件検索によりテキストのカテゴリを特定している。

2

【０００３】

【発明が解決しようとする課題】しかしながら、キーワードパターンは基本的に対象データベースの領域や分野に大きく依存しているばかりでなく、データベースが大規模化するにつれてキーワードの組合せパターンのルールを人手で作成することが困難になってくる。

【０００４】また、分類の判断基準は、カテゴリ特有に作成したキーワードパターンの存在の有無であり、同じキーワードが複数のカテゴリに存在する場合、分類の確からしさを確率的に判断することができなかった。

【０００５】また、対象データベース中のテキストが新聞記事のように具体的な情報の記録や伝達を主目的としている場合は、分類の際に必要になるキーワードには具象物、明確な概念名詞、物理的属性で表現される単語であり、分類のキーワードは主として名詞であるが、手紙文や電報文のように、人間の感覚や感情を伝えることが主目的であるテキストの分類には、形容詞、形容動詞がキーワードとして重要となってくる。

【０００６】このように今後、テキスト自動分類装置の対象データベースの大規模化、広範囲化が進むにつれ、分類ルールを自動的に作成すること、分類ルールに確率的要素を導入することによって、より精度の高い条件判断を行うこと、分類ルールに用いる単語パターンとして名詞だけでなく、形容詞、動詞、形容動詞等の活用する単語も考慮することが新たに必要になる。

【０００７】本発明は、上記に鑑みてなされたもので、その目的とするところは、予めいくつかのカテゴリに分類されたテキストデータベースに対して、カテゴリ固有の単語（名詞、動詞、形容詞、形容動詞）および修飾語・被修飾語対等の言語表現の出現頻度情報を学習することによって新規に入力された日本語テキストを自動的に分類する日本語テキスト自動分類方法を提供することにある。

【０００８】

【課題を解決するための手段】上記目的を達成するため、本発明の日本語テキスト自動分類方法は、日本語のテキストに対して単語および単語の組の頻度を特徴として抽出し、テキストの分類を行う日本語テキスト自動分類方法であって、分類済みテキストアクセス工程にて学習用テキスト蓄積装置に蓄積されている分類ルール抽出のためのテキストをカテゴリ毎にアクセスし、言語表現頻度解析工程にて入力テキスト中の名詞、動詞、形容詞、形容動詞、および修飾語・被修飾語対といった言語表現の出現頻度を計測し、対カテゴリ言語表現頻度テーブル作成工程にて各カテゴリ毎の言語表現の出現頻度の蓄積テーブルを作成し、頻度計測終了判定の後に、対カテゴリ言語表現重要度テーブル作成工程にて、各カテゴリ毎の言語表現の出現頻度を正規化した値の蓄積テーブ

ルを作成する分類ルール自動学習工程と、新規テキスト入力工程にてカテゴリ判定のための新規テキストを入力し、言語表現類似度判定工程にて新規のテキストに出現する言語表現の頻度と、カテゴリ毎の言語表現重要度との類似度を計算した後、該新規テキストのカテゴリを判定し、分類結果出力工程にて前記新規テキストのカテゴリ判定結果を出力するテキスト自動分類工程とを備えたことを要旨とする。

【０００９】また、本発明の日本語テキスト自動分類方法は、前記分類ルール自動学習工程内の言語表現頻度解析において、入力テキストを単語に分割し、名詞、動詞、形容詞、形容動詞といった自立語をラベルし、形態素解析を行う工程と、形態素解析の結果から、修飾語と被修飾語の対を抽出し、修飾語・被修飾語解析を行う工程と、形態素解析と修飾語・被修飾語解析の結果から言語表現のリストを作成する言語表現抽出工程と、入力テキスト中の言語表現の出現頻度を計測する言語表現出現頻度測定工程とを有することを要旨とする。

【００１０】

【作用】本発明の日本語テキスト自動分類方法では、学習用テキスト蓄積装置に蓄積されているテキストをカテゴリ毎にアクセスし、入力テキスト中の名詞、動詞、形容詞、形容動詞および修飾語・被修飾語対といった言語表現の出現頻度を計測し、各カテゴリ毎の言語表現の出現頻度の蓄積テーブルを作成し、各カテゴリ毎の言語表現の出現頻度を正規化した値の蓄積テーブルを作成し、カテゴリ判定のための新規テキストを入力し、新規のテキストに出現する言語表現の頻度とカテゴリ毎の言語表現重要度との類似度を計算した後、該新規テキストのカテゴリを判定し、新規テキストのカテゴリ判定結果を出力する。

【００１１】また、本発明の日本語テキスト自動分類方法では、前記分類ルール自動学習工程内の言語表現頻度解析において、入力テキストを単語に分割し、名詞、動詞、形容詞、形容動詞といった自立語をラベルし、形態素解析を行い、形態素解析の結果から修飾語と被修飾語の対を抽出し、修飾語・被修飾語解析を行い、形態素解析と修飾語・被修飾語解析の結果から言語表現のリストを作成し、入力テキスト中の言語表現の出現頻度を計測する。

【００１２】

【実施例】以下、図面を用いて本発明の実施例を説明する。

【００１３】図１は、本発明の一実施例に係る日本語テキスト自動分類方法を実施する日本語テキスト自動分類装置の構成を示すブロック図である。同図に示す日本語テキスト自動分類装置は、分類ルールの抽出のための学習用テキストを蓄積する学習用テキスト蓄積装置６と、各カテゴリ毎の言語表現の出現頻度を蓄積する対カテゴリ言語表現頻度テーブル７と、各カテゴリ毎の言語表現

の出現頻度を正規化した値を蓄積する対カテゴリ言語表現重要度テーブル８と、前記学習用テキスト蓄積装置６をアクセスして、分類済みのテキストから学習することにより前記対カテゴリ言語表現重要度テーブル７および対カテゴリ言語表現重要度テーブル８を作成する分類ルール自動学習部１７と、カテゴリ判定のための新規のテキストを入力するユーザテキスト入力装置１９と、前記対カテゴリ言語表現重要度テーブル８に蓄積されている分類ルールをアクセスして、前記ユーザテキスト入力装置１９から入力された新規テキストを分類するテキスト自動分類部１８と、該テキスト自動分類部１８で分類された結果を出力表示する分類結果表示装置２０とから構成されている。

【００１４】また、前記分類ルール自動学習部１７は、図２（ａ）に示すように、学習用テキスト蓄積装置６に蓄積されている分類ルール抽出のためのテキストをカテゴリ毎にアクセスする分類済みテキストアクセス部１と、入力テキスト中の名詞、動詞、形容詞、形容動詞、修飾語・被修飾語対といった言語表現の出現頻度を計測する言語表現頻度解析部２と、各カテゴリ毎の言語表現の出現頻度の蓄積テーブルを作成する対カテゴリ言語表現頻度テーブル作成部３と、頻度計測の終点時点を判定する頻度計測終了判定部４と、各カテゴリ毎の言語表現の出現頻度を正規化した値の蓄積テーブルを作成する対カテゴリ言語表現重要度テーブル作成部５とから構成されている。

【００１５】更に、前記テキスト自動分類部１８は、図２（ｂ）に示すように、カテゴリ判定のための新規のテキストを入力する新規テキスト入力部９と、入力テキスト中の名詞、動詞、形容詞、形容動詞、修飾語・被修飾語対といった言語表現の出現頻度を計測する言語表現頻度解析部２と、新規のテキストに出現する言語表現の頻度とカテゴリ毎の言語表現重要度との類似度を計算する言語表現類似度判定部１０と、新規に入力したテキストのカテゴリ判定結果を出力する分類結果出力部１１とから構成されている。

【００１６】また更に、前記言語表現頻度解析部２は、図３に示すように、テキストを入力するテキスト入力部１２と、テキストを単語に分割し、名詞、動詞、形容詞、形容動詞といった自立語をラベルする形態素解析部１３と、形態素解析の結果から、修飾語・被修飾語の対を抽出する修飾語／被修飾語対解析部１４と、形態素解析部１３と修飾語／被修飾語対解析部１４の結果から言語表現のリストを作成する言語表現抽出部１５と、テキスト中の言語表現の出現頻度を計測する言語表現出現頻度測定部１６とから構成されている。

【００１７】以上のように構成される日本語テキスト自動分類装置において、まずテキストを自動的に分類するための分類ルール自動学習部１７について説明する。

【００１８】言語表現とは名詞、動詞、形容詞、形容動

5

詞といった自立語と、自立語の中でも修飾語・被修飾語の関係にある対と定義する。学習用テキスト蓄積装置６には、ｎ個のカテゴリに予め分類されたテキストが蓄積されている。テキストは特に文や章で区切られておらず、同じカテゴリに分類されたテキストが順番に格納されている。分類ルール自動学習部１７の前記分類済みテキストアクセス部１は、c_1 から c_n までのカテゴリに分類されているテキストを順番にアクセスする。ここでは、カテゴリ c_x のテキストを言語表現頻度解析部２に出力する。

【００１９】言語表現頻度解析部２は、図３に示すように、テキスト入力部１２へ入力されたテキストを形態素解析部１３へ出力する。形態素解析部１３では、テキストを形態素解析することにより、単語に分割し、品詞を付与し、リスト形式で出力する。修飾語／被修飾語対解析部１４は、単語に分割されたテキストを解析し、修飾語・被修飾語の関係にある単語の組の対を抽出し、順にリストにして出力する。

【００２０】言語表現抽出部１５では、形態素解析部１３の出力である形態素解析列から名詞、動詞、形容詞、形容動詞といった自立語のみを抽出し、リストを作成する。また、修飾語／被修飾語対解析部１４の出力である修飾語・被修飾語の対のリストも結合し、言語出現頻度測定部１６へ出力する。

【００２１】言語出現頻度測定部１６では、言語表現の出現頻度を測定する。カテゴリ c_x のテキストに対して、言語表現 t_k が出現した頻度 d_{xk} をカウントし、図４の対カテゴリ言語表現頻度テーブル上の c_x の列に格納する。そして、頻度計測終了判定部４において $x=n$ になるまで、この作用を全てのカテゴリに対して繰り返し、対カテゴリ言語表現頻度テーブル７を作成する。

【００２２】対カテゴリ言語表現重要度テーブル作成部５は、対カテゴリ言語表現頻度テーブル７を正規化する。正規化の計算式は

【数１】

$$w_{ij} = \frac{d_{ij}}{\sum_{i=1}^{N} d_{ij}}$$

$$S_i\ (V_i,\ N) = \frac{\sum_{k=1}^{n} (w_{ik} \cdot y_k)}{\sum_{k=1}^{n} w_{ik} + \sum_{k=1}^{n} y_k - \sum_{k=1}^{n} (w_{ik} \cdot y_k)}$$

この類似度 S_i を全てのカテゴリについて計算する。類似度 S_i が $i=1$ で最大となった場合、新規入力テキストのカテゴリは c_1 と判定される。

【００２８】次に具体例として、予め分類されている電報文データベースを用いた学習と新規に入力された電報文のカテゴリの判定例を説明する。

【００２９】学習用テキスト蓄積装置６には、電報文が結婚式、結婚記念日、誕生日、卒業式といった目的に応じたカテゴリに分類・蓄積されている。まず、カテゴリ「結婚式」に分類されている電報文例の学習について説

6

とする。ここで、d_{ij} はカテゴリ c_i のテキスト中に存在した言語表現 t_j の頻度、w_{ij} はカテゴリ c_i に対する言語表現 t_j の重要度である。w_{ij} は言語表現 t_j がある特定のカテゴリ c_i 中にどれだけの割合で存在していたかを示す。この w_{ij} を全ての t_{ij} に対して求め、図５に示す対カテゴリ言語表現重要度テーブル８を作成する。

【００２３】次に、テキスト自動分類部１８について説明する。

【００２４】新規のテキストをテキスト自動分類部１８の前記新規テキスト入力部９に入力すると、テキストは言語表現頻度解析部２に入力される。言語表現頻度解析部２では、入力されたテキストを形態素解析、構文解析を行った後、自立語、修飾語・被修飾語の対を抽出し、対カテゴリ言語表現重要度テーブル８上の言語表現 t_k の新規テキスト中における出現頻度をカウントする（図６）。この新規テキストにおける出現頻度を１次元配列で表現すると、

$N = (y_1, y_2, \cdots, y_n)$

ここで、y_j は新規例文中の言語表現 t_j の出現頻度である。

【００２５】作成された新規テキスト言語表現頻度分布は言語表現類似度判定部１０に入力され、対カテゴリ言語表現重要度テーブル８の各カテゴリ c_i 毎に類似度 S_i を計算する。カテゴリ c_i に対する言語表現 t の頻度を１次元配列で表現すると、

$V_i = (w_{i1}, w_{i2}, \cdots, w_{in})$

ここで、w_{in} は言語表現 t_n のカテゴリ c_i の頻度である。

【００２６】新規テキストがこのカテゴリ c_i に属する確からしさは１次元配列Ｎと V_i の類似度 S_i で表現する。

【００２７】

【数２】

明する。「結婚式」例文データベース中に存在する言語表現の頻度を測定する。分類済みテキストアクセス部１が学習用テキスト蓄積装置６のカテゴリc_1「結婚式」の第一文にアクセスする。

【００３０】

”春の微風に乗って、新しい門出おめでとう。二人仲良く、めざせ２１世紀”

この文は言語表現頻度解析部２へ入力される。言語表現頻度解析部２のテキスト入力部１２に入力された電報文は、形態素解析部１３において形態素解析され、名詞、動詞、形容詞、形容動詞は自立語とマークされる。この例では次のように形態素に分割される。”／”は形態素の区切り記号である。

”春（自立語）／の／微風（自立語）／に
／乗っ（自立語）／て／、／新しい（自立語）
／門出（自立語）／おめでとう（自立語）／。
／二人（自立語）／仲良く（自立語）／、
／めざせ（自立語）／２１世紀（自立語）”

形態素が動詞・形容詞・形容動詞の場合には終止形情報も付与する。言語表現抽出部１５では、自立語と修飾語・被修飾語の対を抽出する。自立語は終止形で抽出される。

【００３１】

（春　微風　乗る　新しい　門出　おめでとう
二人　仲良い　めざす　２１世紀）

修飾語／被修飾語対解析部１４では修飾語・被修飾語の関係にある自立語の対を抽出する。

【００３２】

((春　微風）（新しい　門出）（二人　仲良い))

言語表現抽出部１５は最終的に自立語と、修飾語・被修飾語のリストを結合し、出力する。

【００３３】

（春　微風　乗る　新しい　門出　おめでとう
二人　仲良い　めざす　２１世紀
（春　微風）（新しい　門出）（二人　仲良い））

言語表現出現頻度測定部１６は、１つの言語表現に対して、対カテゴリ言語表現頻度テーブル７中のc_1の列に頻度を記憶する変数を確保し、頻度を書き込む。この最初の例文の場合はどの言語表現も１度しか出現していないので、頻度は１となる。この作用をカテゴリc_1「結婚式」の全ての電報例文について行い、カテゴリc_1「結婚式」中に存在する言語表現tと、その頻度の１次元配列が対カテゴリ言語表現頻度テーブル７に書き込まれる。

【００３４】頻度計測終了判定部４から、再び分類済みテキストアクセス部１へ戻り、分類済みテキストアクセス部１ではカテゴリc_2のテキストについて、上述したと同じ作用を繰り返し行う。頻度計測が全てのカテゴリについて終わると、対カテゴリ言語表現頻度テーブル７（図７）が完成する。

【００３５】対カテゴリ言語表現重要度テーブル作成部５では、対カテゴリ言語表現頻度テーブル７を参照し、対カテゴリ言語表現重要度テーブル８を作成する。例えば言語表現t_1「おめでとう」のカテゴリc_1「結婚式」における、重要度w_{11}を計算すると、

【数３】

$$w_{11}=\frac{d_{11}}{\sum_{i=1}^{N} d_{i1}}=\frac{130}{130+120+100+96+0}=0.29$$

となる。この重要度を５つのカテゴリc、全ての言語表現tについて求め、図８に示す対カテゴリ言語表現重要度テーブル８に書き込む。

【００３６】次に、判定部の具体的な例を次の電報例文で説明する。

「ご結婚おめでとう。二人で植えよう愛の木を。
そして咲かせよう、幸せの花を。」

新規テキスト入力部９に入力された電報例文は、言語表現頻度解析部２へ出力され、言語表現tが抽出される。

【００３７】

（結婚　おめでとう　ふたり　植える　愛　木
咲く　幸せ　花　（愛　木）（幸せ　花））

抽出された言語表現の頻度分布を図９に示す。例では、全ての言語表現について頻度を図示することができないので、この表の範囲の言語表現だけで、カテゴリ「結婚式」における類似度を計算する。新規テキストにおける出現頻度を１次元配列で表現すると、

【数４】Ｎ＝（１，０，０，０，０，０，０，０，１，１，０，０）

ここで対カテゴリ言語表現重要度テーブル８を参照し、カテゴリ「結婚式」における言語表現の重要度を１次元配列にすると、

【数５】V_1＝（０．２９，０．７１，０．５３，０．２２，０．５０，０．４５，０．１４，０．２７，０．４９，０．７９，０．２３，０．００）

類似度S_1を計算すると、

【数６】

$$S_1(V_1, N) = \frac{\sum_{k=1}^{n}(w_{1k} \cdot y_k)}{\sum_{k=1}^{n} w_{1k} + \sum_{k=1}^{n} y_k - \sum_{k=1}^{n}(w_{1k} \cdot y_k)} = 0.26$$

同様の計算を他のカテゴリについて行うと、カテゴリｉに対する類似度は、

【数７】S_n ＝（０．２６，０．０９，０．０８，０．１５，０．００）

と求められ、類似度が最大となるのはS_1 ＝０．２６の場合であり、対応するカテゴリc_1「結婚式」の電報文と分類される。

【００３８】上述したように、本発明の日本語テキスト自動分類方法は、言語表現の頻度の測定対象として名詞だけでなく、動詞、形容詞、形容動詞等の活用する単語、修飾語・被修飾語の関係にある単語の対も対象にしている点、予め分類されたテキスト中の言語表現の頻度から各カテゴリに対する言語表現重要度テーブルを作成する点、および新規に入力されたテキストの言語表現出現頻度を測定し、対カテゴリ言語表現重要度テーブルとの類似度を計算することによって入力テキストを分類する点に特徴があり、従来の技術と異なる。

【００３９】

【発明の効果】以上説明したように、本発明によれば、テキスト分類のためのルールを人手で作成することなしに、カテゴリ特有に出現する名詞のみならず、形容詞、動詞、形容動詞や修飾語・被修飾語の対といった言語表現の頻度のパターンを自動的に抽出し、新規に入力されるテキストを言語表現の頻度パターンとの類似度を計算することによって、最も確からしいカテゴリにテキストを分類することができる。

【図面の簡単な説明】

【図１】本発明の一実施例に係る日本語テキスト自動分類方法を実施する日本語テキスト自動分類装置の構成を示すブロック図である。

【図２】図１の日本語テキスト自動分類装置に使用されている分類ルール自動学習部およびテキスト自動分類部の構成を示すブロック図である。

【図３】図２に示す分類ルール自動学習部およびテキスト自動分類部に使用されている言語表現頻度解析部の構成を示すブロック図である。

【図４】図１の日本語テキスト自動分類装置に使用されている対カテゴリ言語表現重要度テーブルを示す図である。

【図５】図１の日本語テキスト自動分類装置に使用されている対カテゴリ言語表現重要度テーブルを示す図である。

【図６】新規テキストの言語表現出現頻度テーブルを示す図である。

【図７】対カテゴリ言語表現頻度テーブルの一例を示す図である。

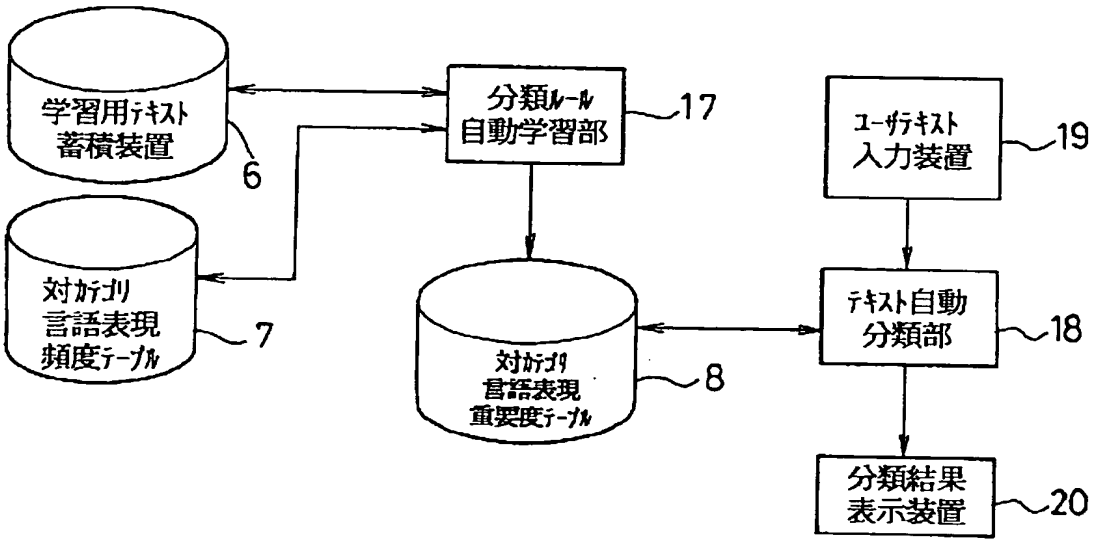
【図８】対カテゴリ言語表現重要度テーブルの一例を示す図である。

【図９】新規テキストの言語表現出現頻度テーブルの一例を示す図である。

【符号の説明】

１　分類済みテキストアクセス部
２　言語表現頻度解析部
３　対カテゴリ言語表現頻度テーブル作成部
４　頻度計測終了判定部
５　対カテゴリ言語表現重要度テーブル作成部
６　学習用テキスト蓄積装置
７，８　対カテゴリ言語表現重要度テーブル
９　新規テキスト入力部
１０　言語表現類似度判定部
１３　形態素解析部
１４　修飾語／被修飾語対解析部
１５　言語表現抽出部
１６　言語表現出現頻度測定部
１７　分類ルール自動学習部
１８　テキスト自動分類部
１９　ユーザテキスト入力装置

【図１】

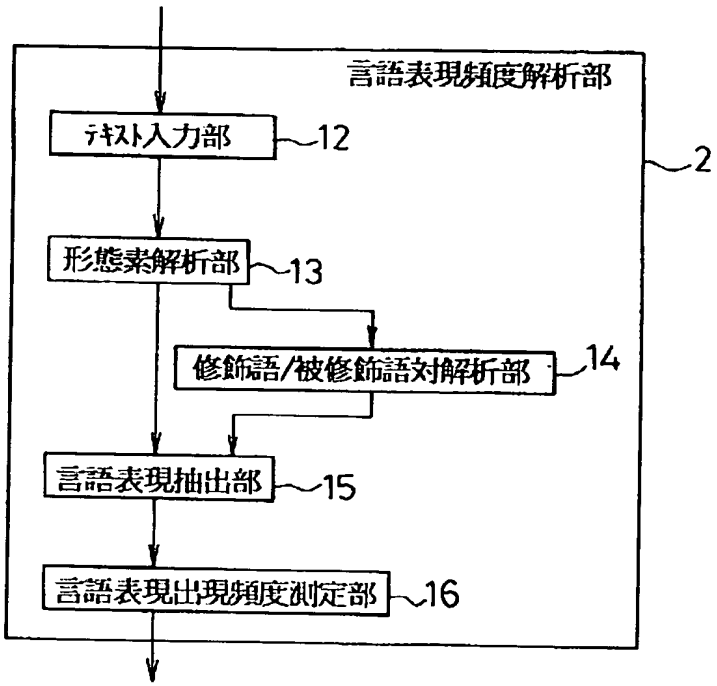


【図６】

新規ﾃｷｽﾄ中の言語表現 t_k の出現頻度

言語表現 t_k	
t_1	y_1
t_2	y_2
t_3	y_3
⋮	⋮
t_k	y_k

【図３】



【図４】

分類ｶﾃｺﾞﾘ c_n

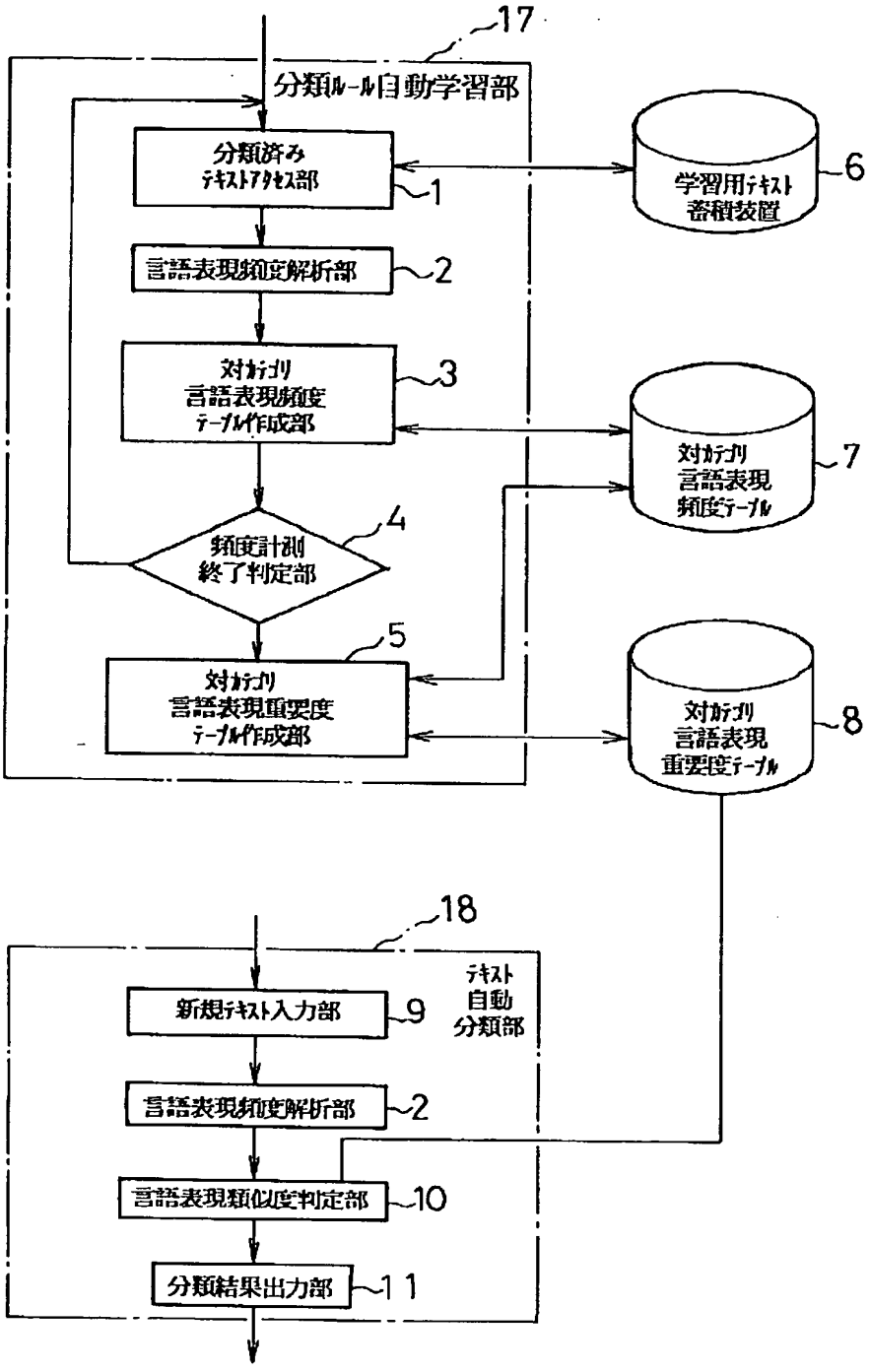
言語表現 t_k	c_1	c_2	c_3	・・・	c_n
t_1	d_{11}	d_{21}	d_{31}	・・・	d_{n1}
t_2	d_{12}	d_{22}	d_{32}	・・・	d_{n2}
t_3	d_{13}	d_{23}	d_{33}	・・・	d_{n3}
⋮	⋮	⋮	⋮	・	⋮
t_k	d_{1k}	d_{2k}	d_{3k}	・・・	d_{nk}

【図５】

分類ｶﾃｺﾞﾘ c_n

言語表現 t_k	c_1	c_2	c_3	・・・	c_n
t_1	w_{11}	w_{21}	w_{31}	・・・	w_{n1}
t_2	w_{12}	w_{22}	w_{32}	・・・	w_{n2}
t_3	w_{13}	w_{23}	w_{33}	・・・	w_{n3}
⋮	⋮	⋮	⋮	・	⋮
t_k	w_{1k}	w_{2k}	w_{3k}	・・・	w_{nk}

【図２】



【図９】

新規テキスト中の
言語表現ｉ出現頻度

言語表現ｉ

おめでとう	1
永遠	0
仲良く	0
頑張る	0
乾杯	0
信じる	0
思い出	0
涙	0
幸せ	1
結婚	1
：	：
（新しい 門出）	0
（強い 子）	0

【図７】

分類析月 c_n

言語表現 t_n	結婚式	誕生日	卒業式	出産	お悔やみ
おめでとう	130	120	100	96	0
永遠	96	3	5	2	30
仲良く	40	20	13	3	0
頑張る	35	15	65	45	2
乾杯	53	30	15	5	2
信じる	30	5	4	3	25
思い出	15	10	60	5	20
涙	25	9	15	10	35
幸せ	80	16	26	40	0
結婚	150	2	3	34	0
：	：	：	：	：	：
(新しい 門出)	7	0	24	0	0
(強い 子)	0	27	3	47	0

【図８】

分類析月 c_n

言語表現 t_n	結婚式	誕生日	卒業式	出産	お悔やみ
おめでとう	0.29	0.27	0.22	0.22	0.00
永遠	0.71	0.02	0.04	0.01	0.22
仲良く	0.53	0.26	0.17	0.04	0.00
頑張る	0.22	0.09	0.40	0.28	0.01
乾杯	0.50	0.29	0.14	0.05	0.02
信じる	0.45	0.07	0.06	0.04	0.37
思い出	0.14	0.09	0.55	0.05	0.18
涙	0.27	0.10	0.16	0.11	0.37
幸せ	0.49	0.10	0.16	0.25	0.00
結婚	0.79	0.01	0.02	0.18	0.00
：	：	：	：	：	：
(新しい 門出)	0.28	0.00	0.77	0.00	0.00
(強い 子)	0.00	0.35	0.04	0.61	0.00